LAMBDA @ HSE CS

LAboratory for Methods for Big Data Analysis

August, 2020

Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

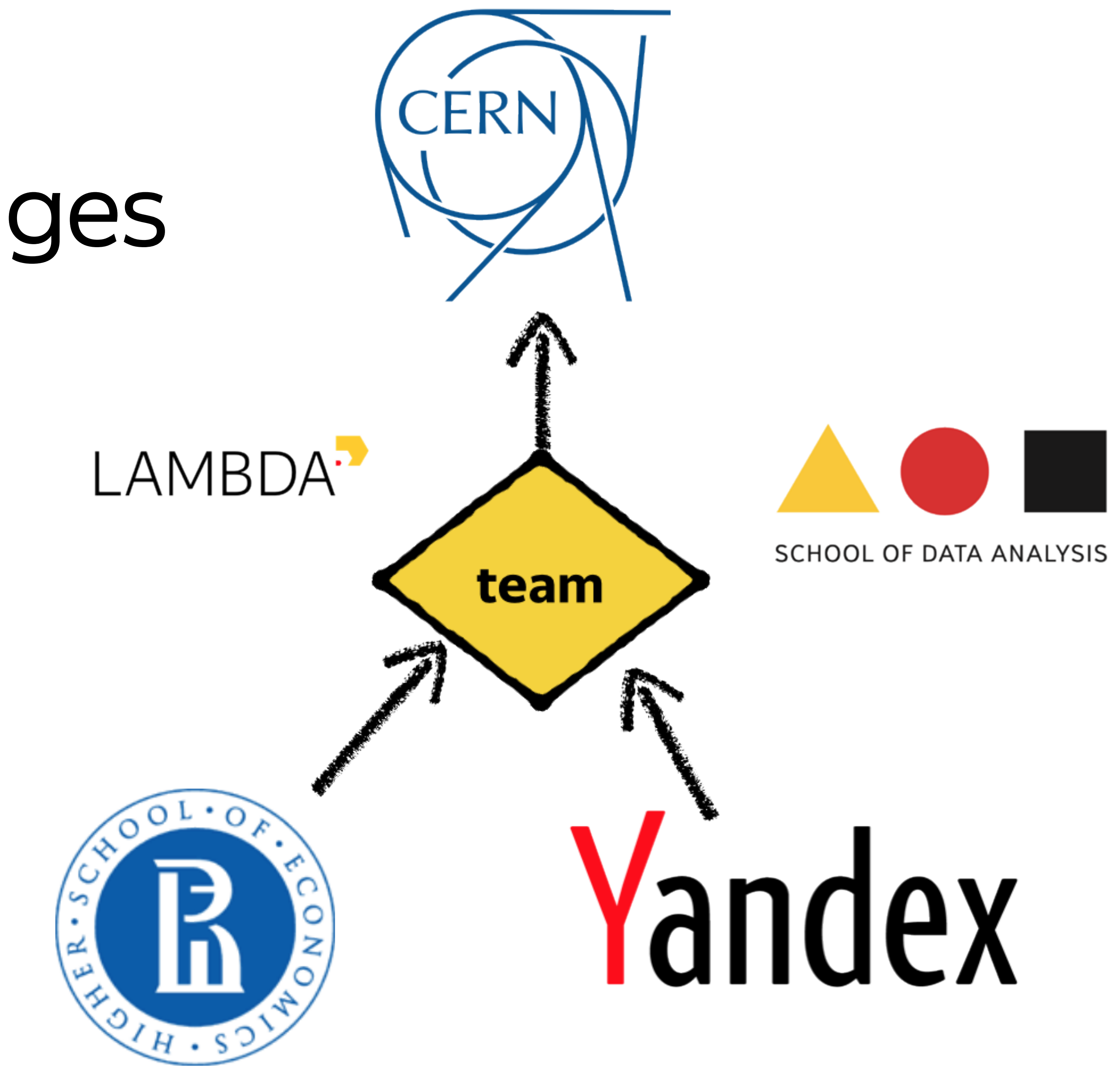
Quick self-intro

Head of LHCb Yandex School of Data Analysis (YSDA) team
Head of Laboratory (link) of methods for Big Data Analysis at Higher School of Economics (HSE),

- Applications of Machine Learning to **natural science challenges**
- HSE has joined LHCb in 2018!

Co-organizer of Flavours of Physics @Kaggle (2015)
Co-organizer of TrackML challenge (2018)
Education activities (ML at ICL, ClermonFerrand, URL Barcelona, Coursera)

- Summer school on Machine Learning
  in Hamburg, 2019, Oxford 2018, Reading 2017, Lund 2016, …



Main Laboratory Focus

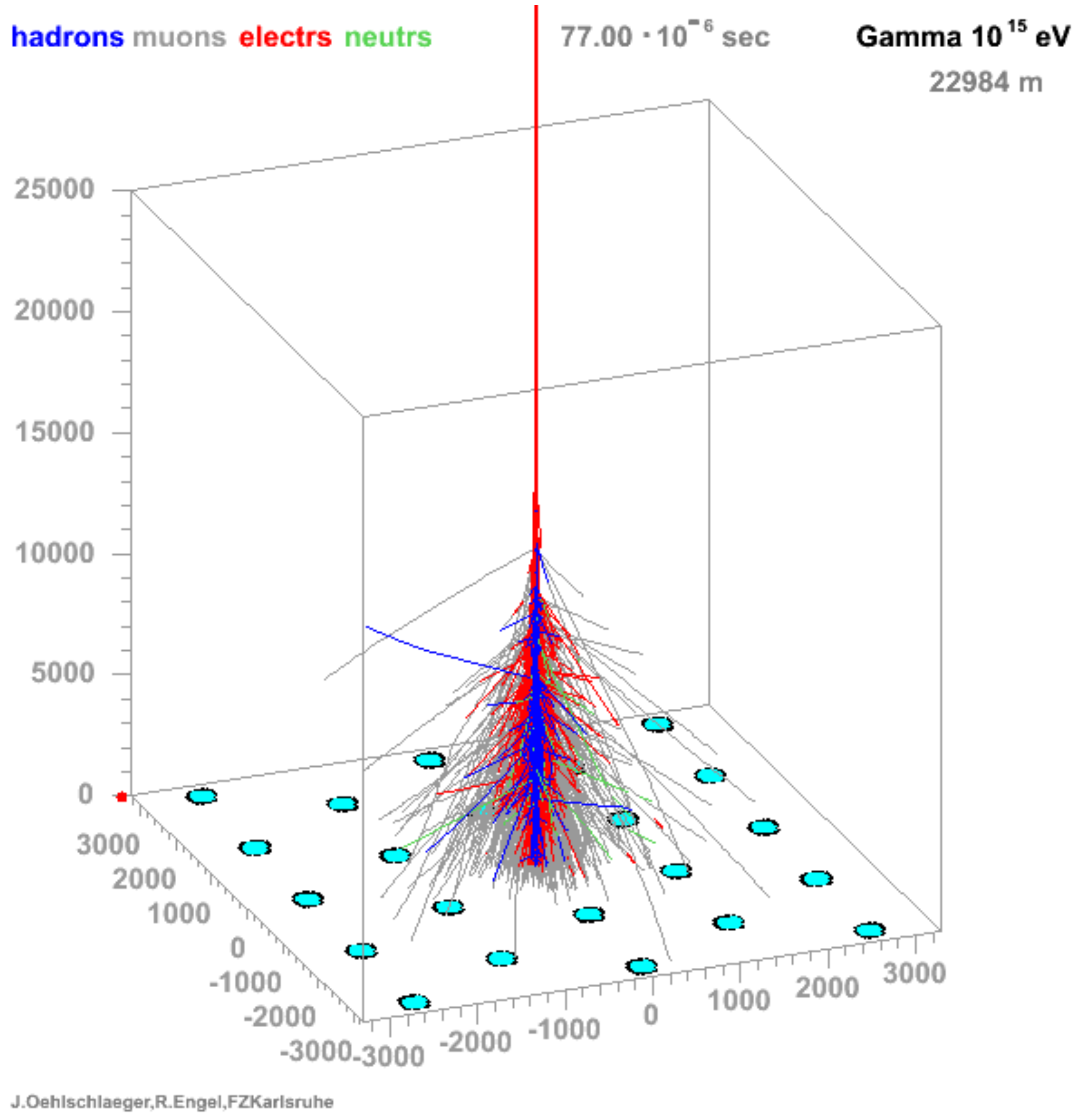
Development of Machine Learning methods for solving tough fundamental science challenges;

Collaboration with leading international research institutions and industry for solving advanced applied problems;

Promoting Machine Learning in Natural Science communities.

Scientific Research Highlights

CRAYFIS muon trigger for



Up to

98%

speedup for running deep neural net model

Task

CRAYFIS experiment proposes usage of private mobile phones for observing Ultra-High Energy Cosmic Rays. Distributed observatory, seeking for particles of energies > 10^{18} eV. Design trigger for mobile device that can catch

- an intensive air shower from UHECR (occurs in less than microseconds);
- supports high frame rate (10 Hz)
- trigger on minimally ionizing particles (assuming that such particles leave traces with brightness comparable to the level of intrinsic camera noise).

Data used

- CRAYFIS Simulated sample

ML Metrics

Linear combination of

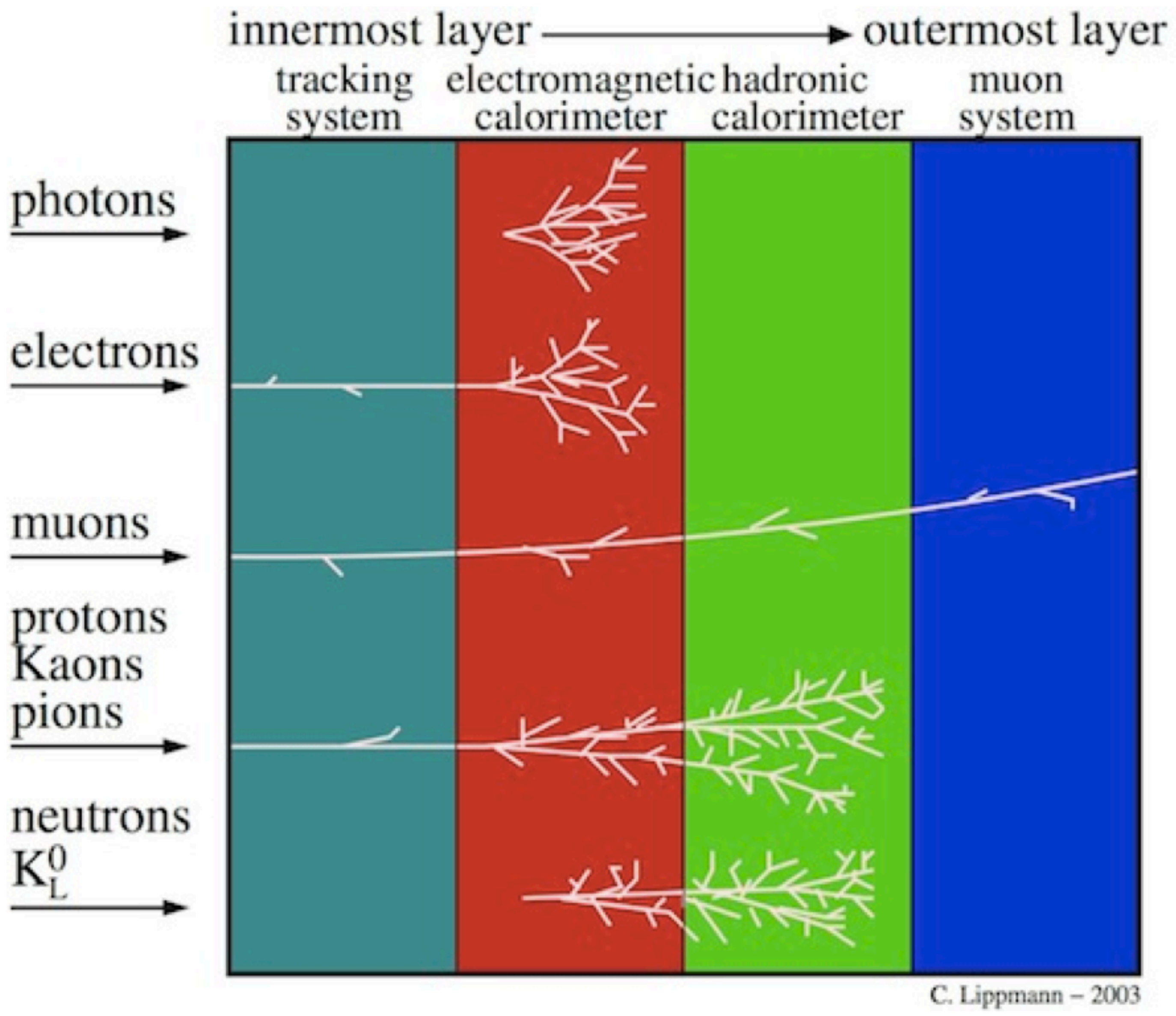
- weighted cross-entropy;
- computational complexity.

Result

- for just 1.4 times more computational cost than simple cut, gives signal efficiency of 90% and background rejection 60%;
- computational complexity is 0.02 of regular convolutional network;
- http://bit.ly/2nb7gfx

LHCb particle identification

Case



Up to

50%

algorithm error reduction

Task

identify charged particle associated with a track (multiclass classification problem);

particle types: Electron, Muon, Pion, Kaon, Proton and “Ghost”;

combine information from LHCb subdetectors: CALO, RICH, Muon and Tracker;

Data used

- LHCb Simulated sample

ML Metrics

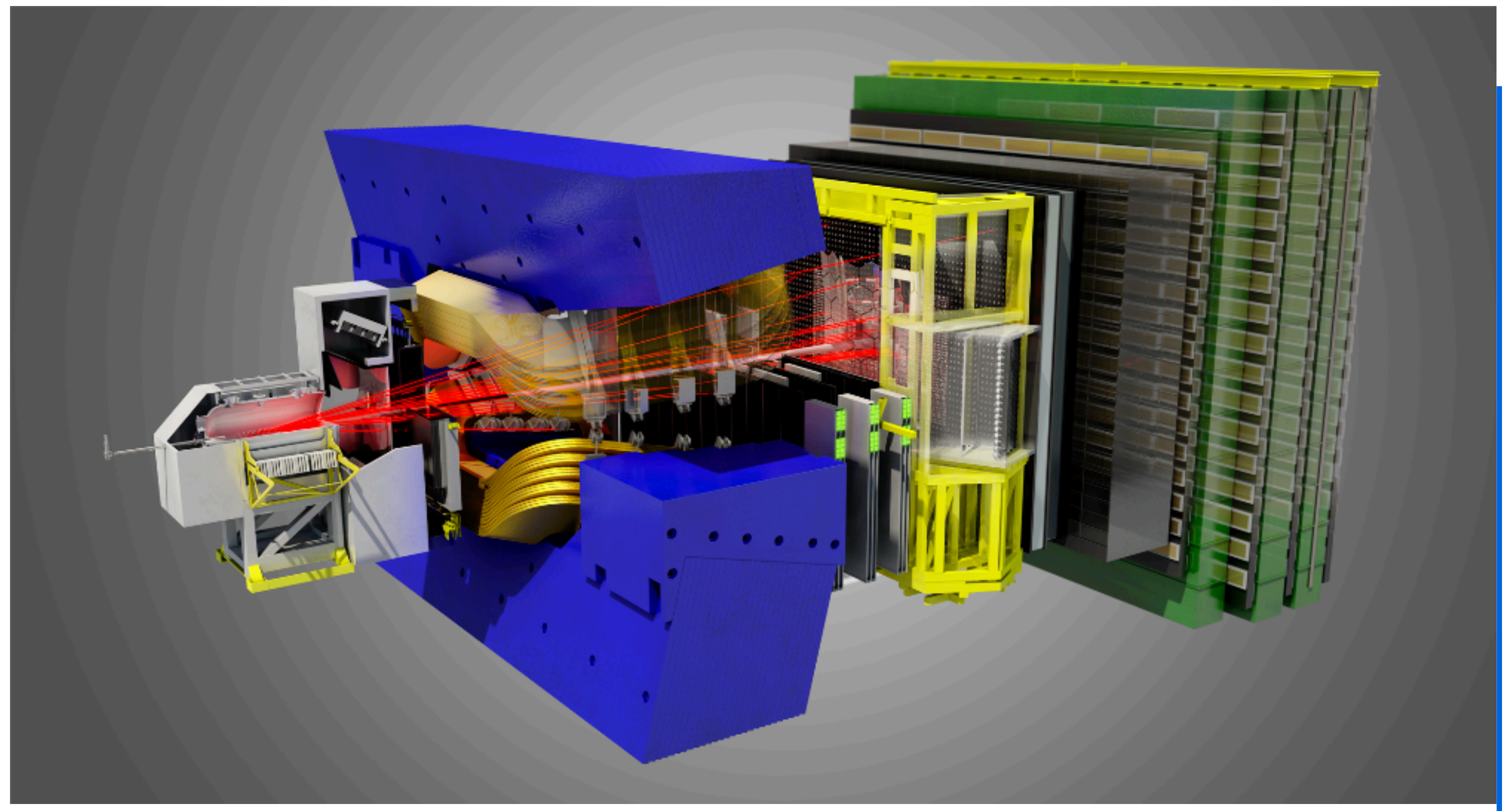
- ROC AUC one vs all,
- model flatness

Result

- Blended NN model that has error rate half less than baseline for some of the particles;
- Blended BDT model with same ROC AUC, but that is flat wrt given features;
- http://bit.ly/2l0yvXc

LHCb particle ID generation

Case



Neural Net-based generator simulator is

100x

faster than the full simulation

Task

generate particle identification probability for simulated samples

particle types: Electron, Muon, Pion, Kaon, Proton and “Ghost”

based on information from LHCb subdetectors

Data used

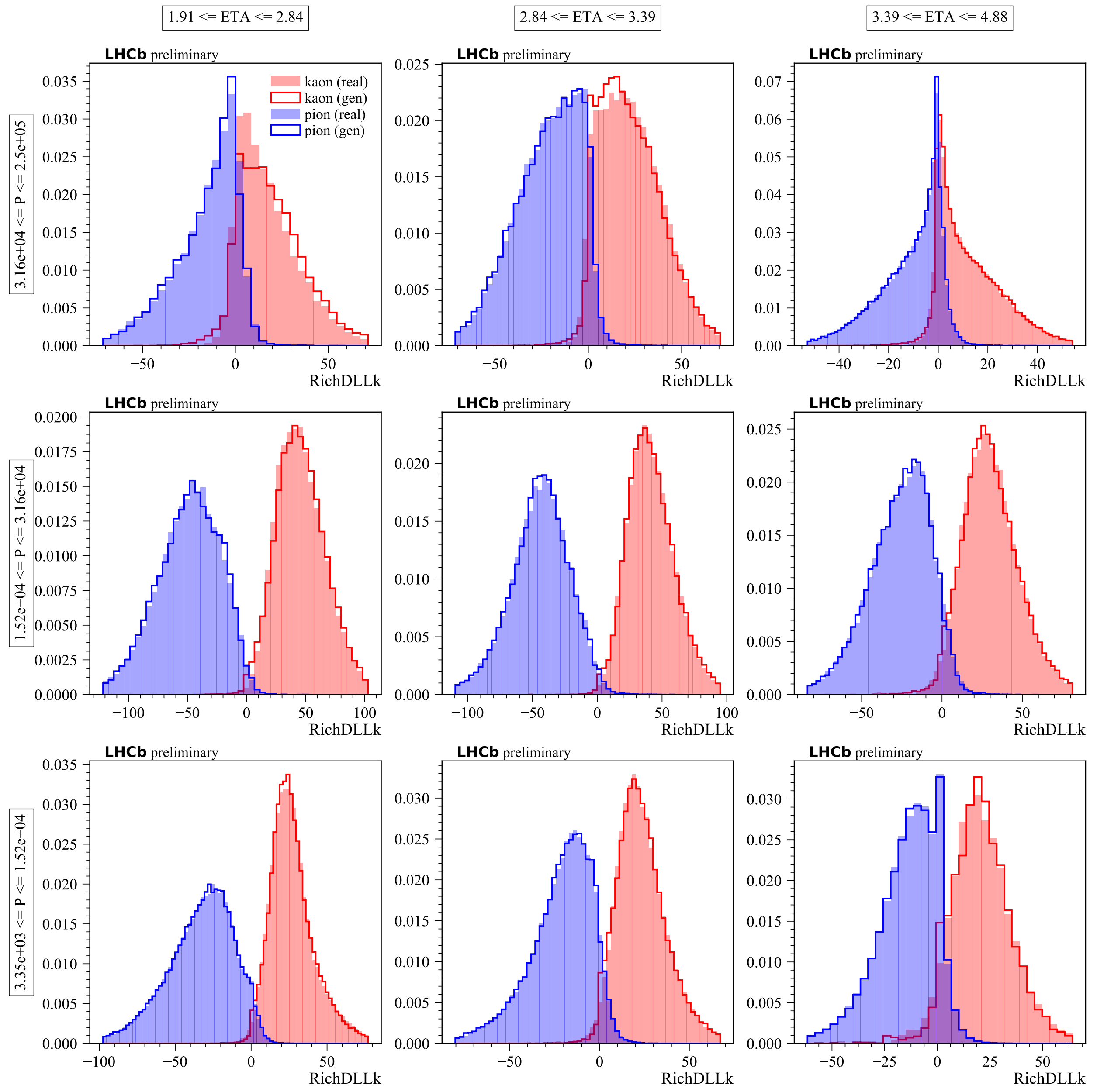
- LHCb real calibration samples

ML Metrics

- AUC similarity

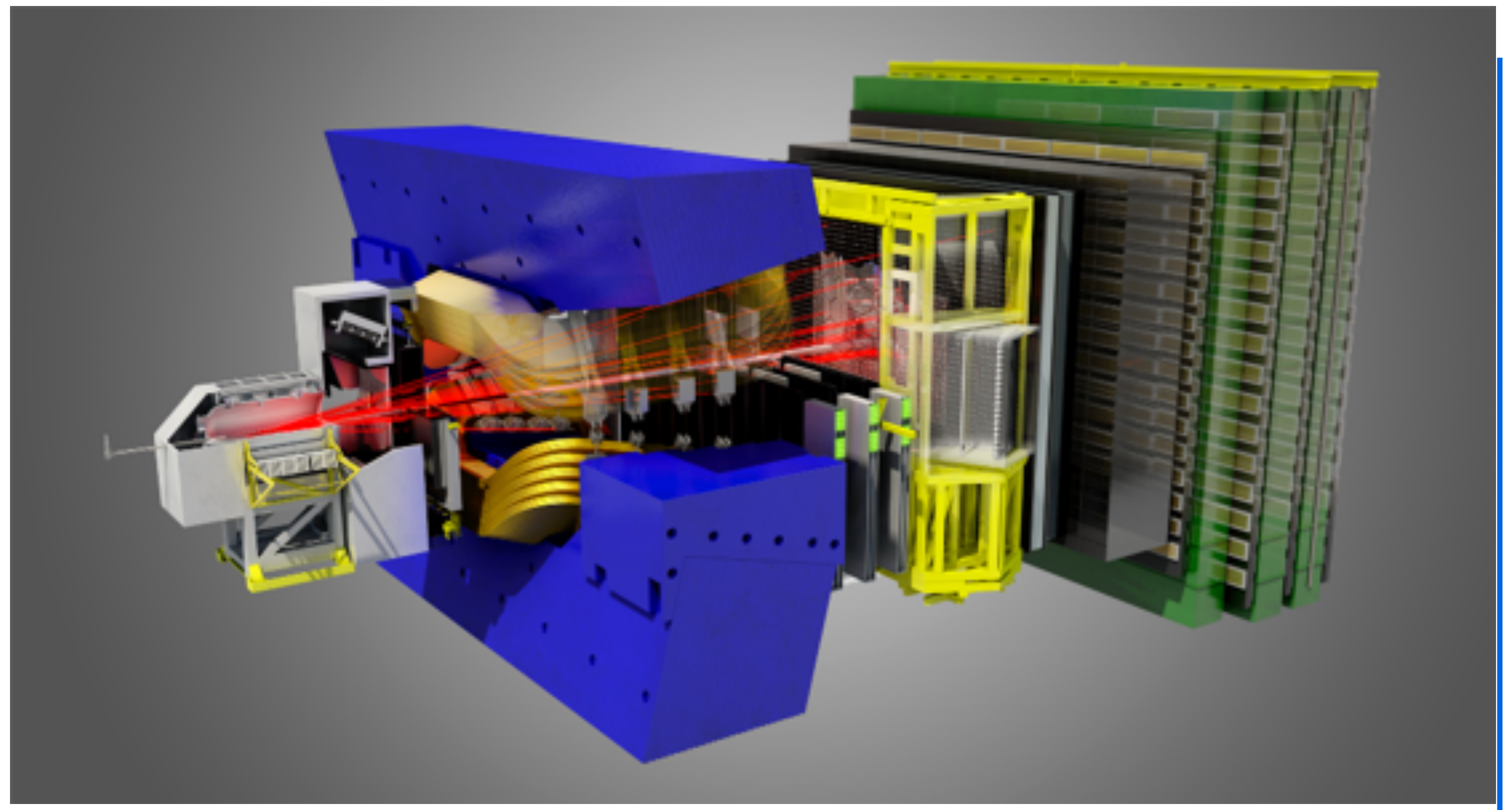
Result

- NeuralNet-based simulator is 100 times faster than the full simulation
- being trained on real data, emulator is more precise then the one based on full simulation
- https://doi.org/10.1016/j.nima.2019.01.031



LHCb fast simulation of detector response

Case



Task

generate stochastic response in the calorimeter of the detector

particle types: Electron, Photon, Pion, Kaon, Proton

Data used

› LHCb simulated samples

ML Metrics

› physics properties of generated clusters

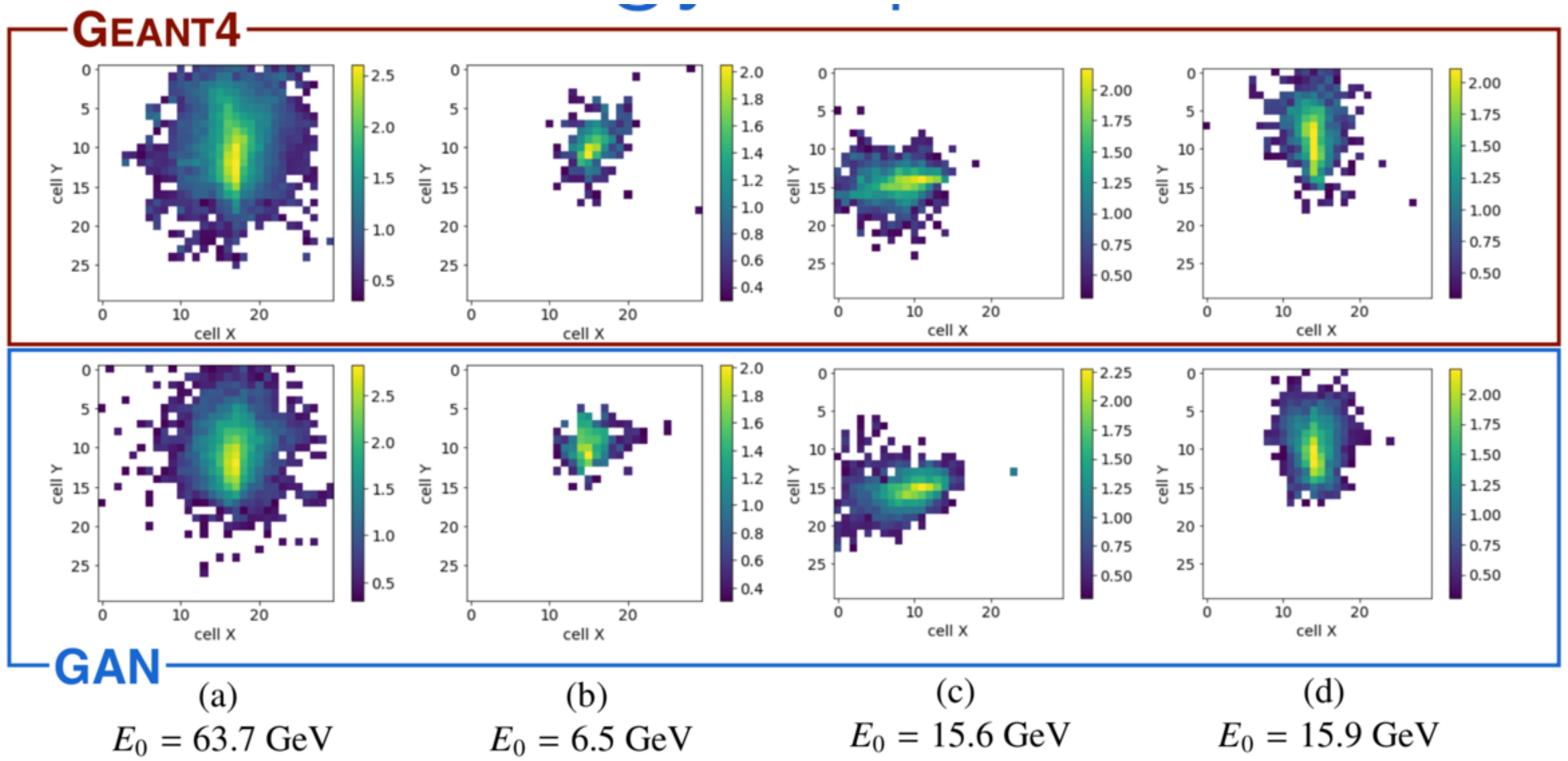
Result

› NeuralNet-based simulator is 1000 times faster than the full simulation

› https://doi.org/10.1051/epjconf/201921402034

1000x

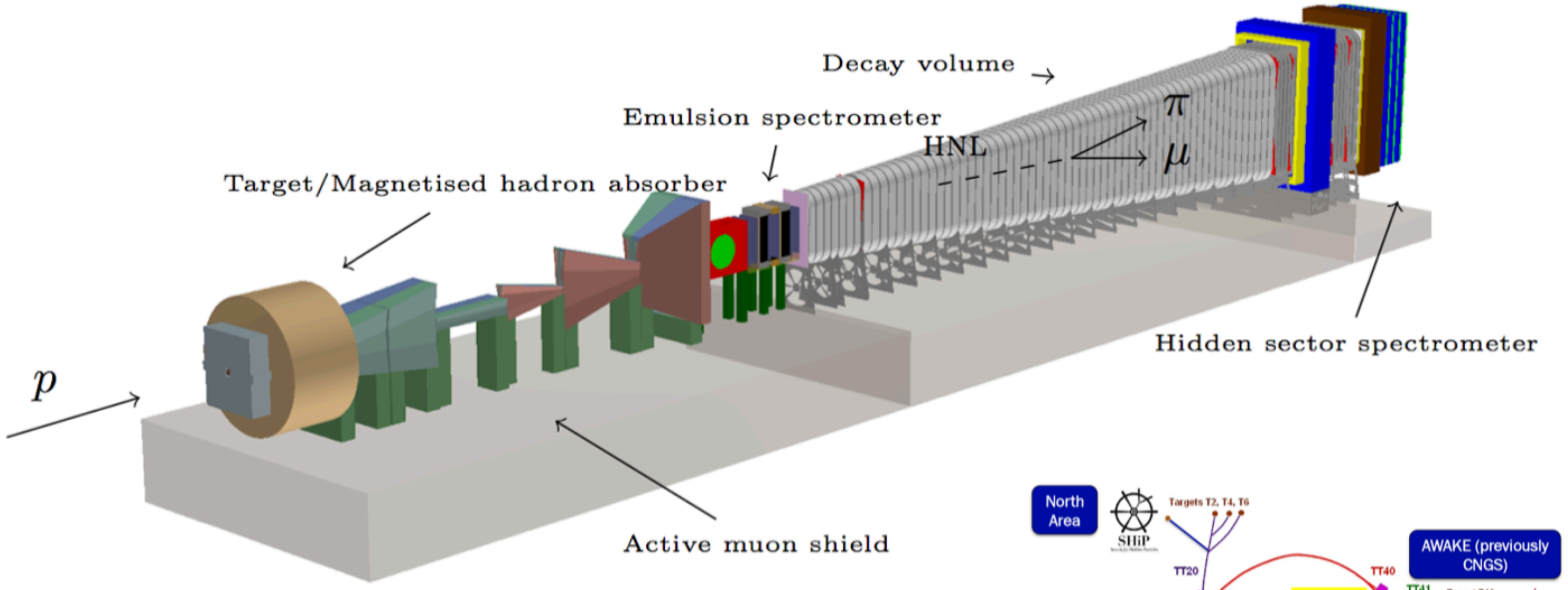
speed-up



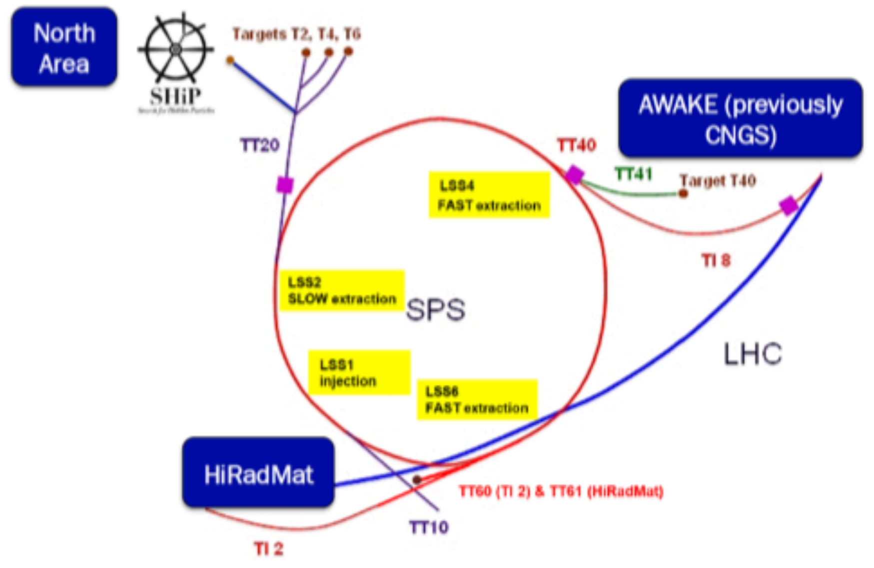
SHiP shield optimization

Problem: find complex shape of muon deflection component. 50-dimensional space.

Solution: with the help of non-gradient optimization and advanced simulation technique we have discovered configuration that is 25% lighter (saves CHF 1M) and has the same physics properties.



◇ Search for Hidden Particles



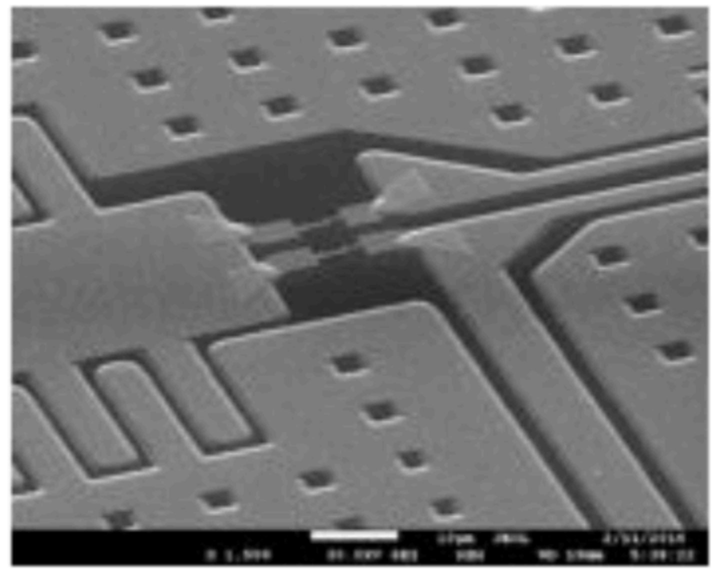
http://iopscience.iop.org/article/10.1088/1742-6596/934/1/012050/meta

Quantum system control

Problem: learn to control qbit to switch from one state to another

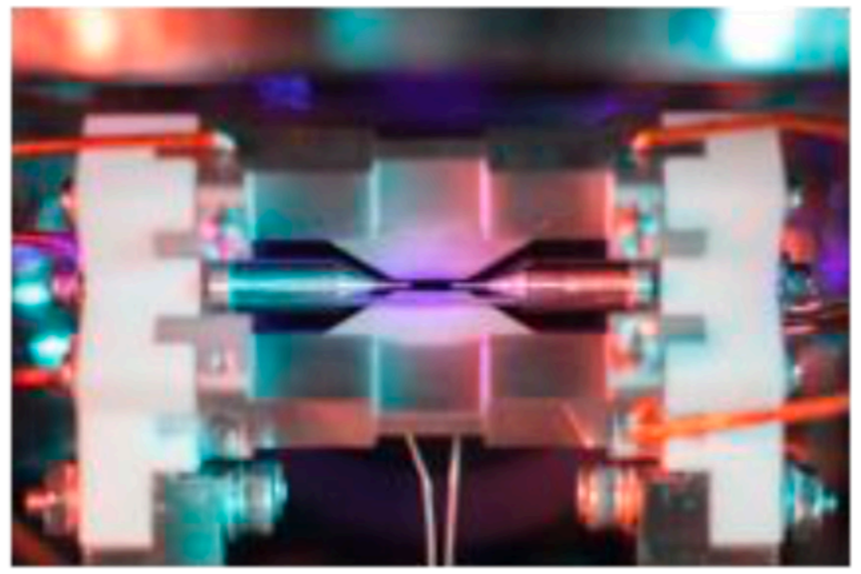
Solution: with the help of Reinforcement Learning and differentiable simulator the accuracy and speed of convergence has increased dramatically.

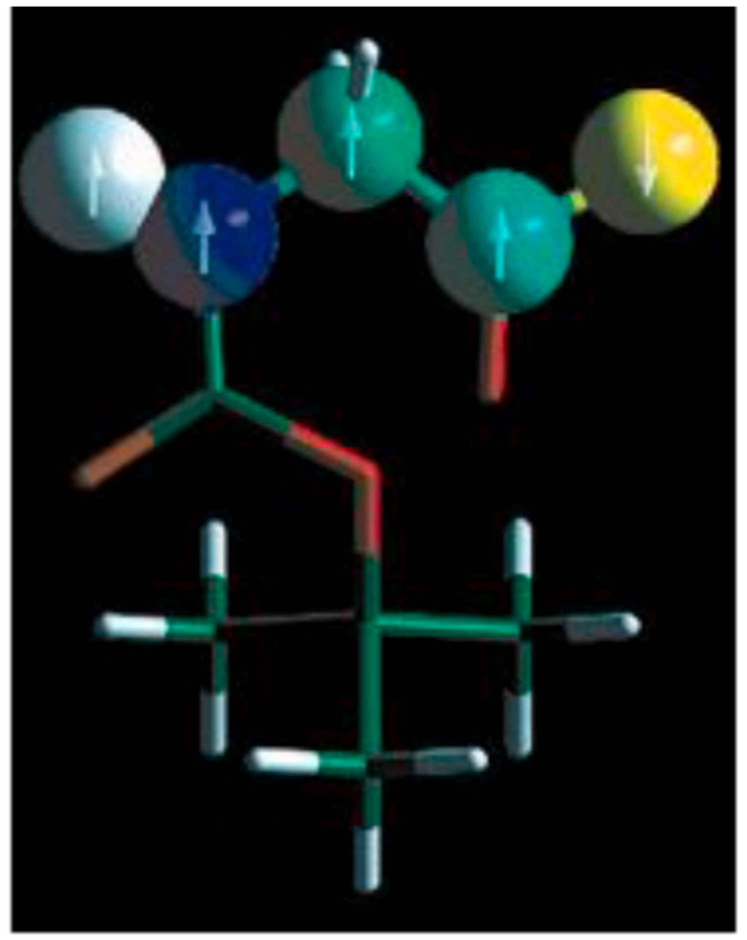
Status: waiting for experimental confirmation of those results



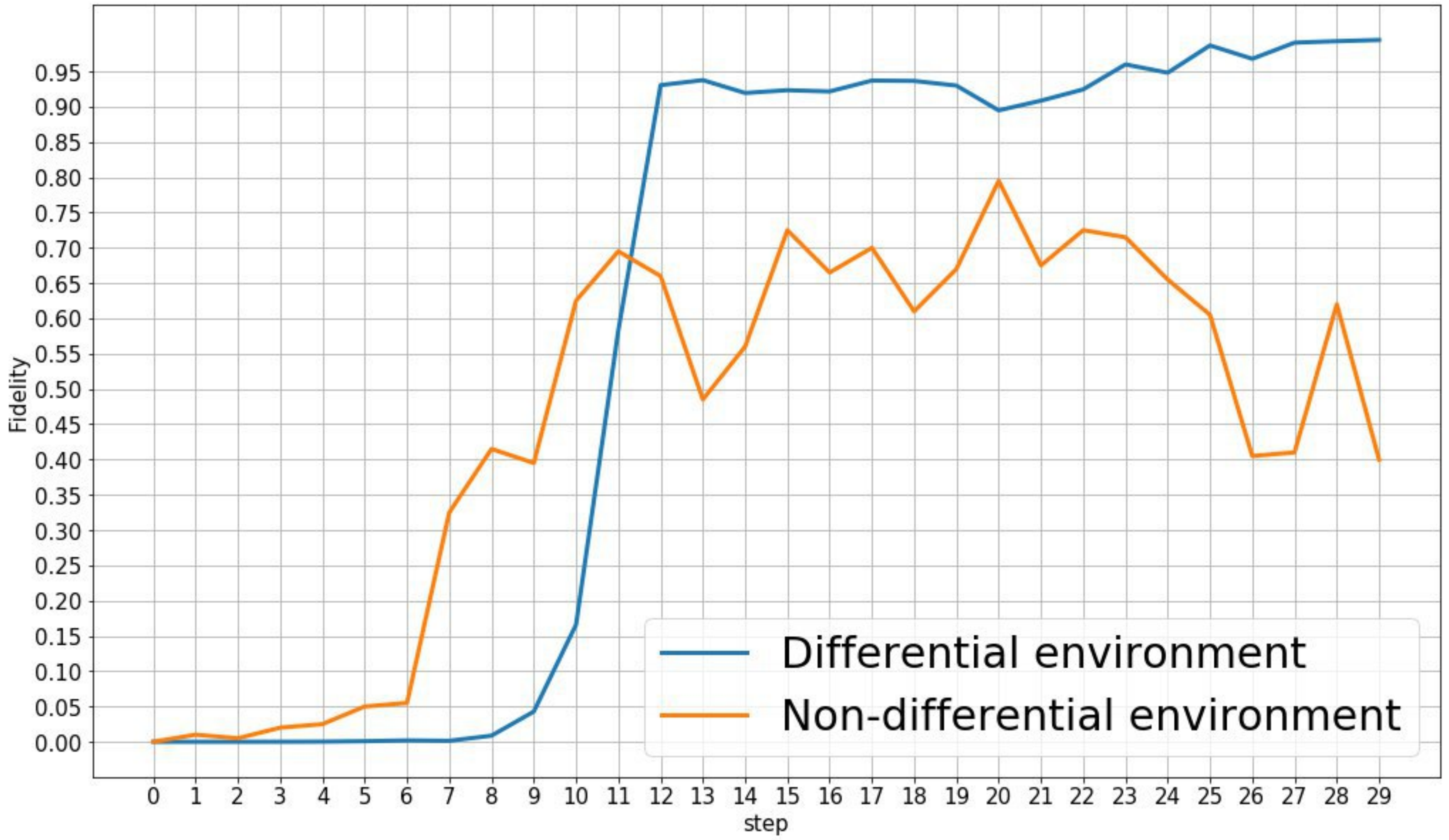
Superconducting quantum circuit

Atomic traps





NMR



“Industrial” Research Highlights

Data storage optimisation

Client case

CERN openlab

Challenge

The Large Hadron Collider detectors take captures of every notable particle interaction, which piles up to over 5 PB of data a year. But the market price for 1 PB of data storage per year runs as high as 1 million dollars.

$4m

maximum yearly projected savings on data storage

Task

To cut storage costs and to determine which files should be stored on which kind of medium, to improve the effectiveness of data access

Data used

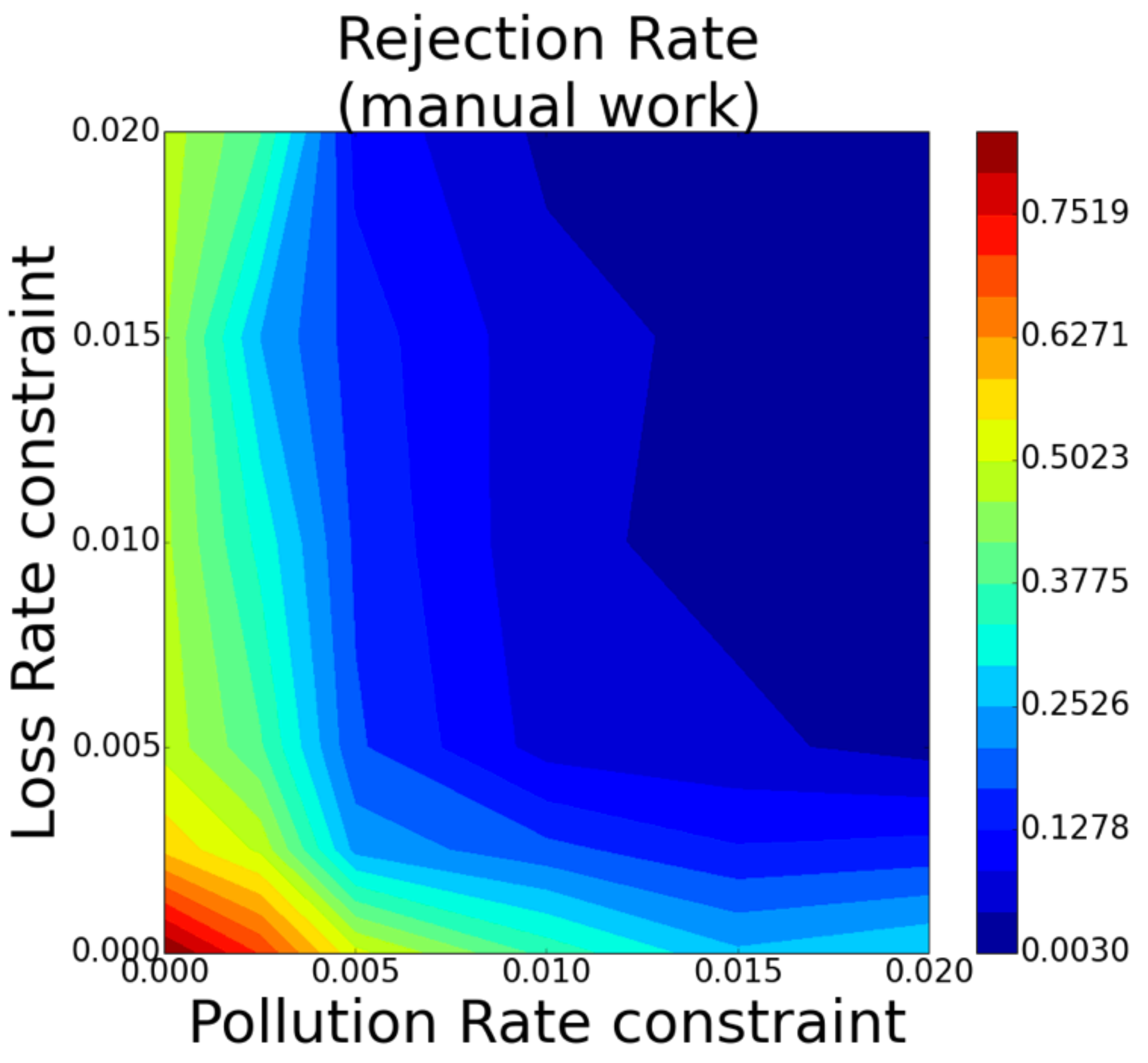
- Historical data on the access history of every file generated by LHCb and the collision simulators (each file catalogued by several features, like file size, number of existing file copies, access frequency, longest duration for which the files hadn't been accessed, file origin, etc.)

Result

- Data storage optimisation by 40%
- A model that allows saving up to 4 petabytes (more than 4 million gigabytes) of storage a year – the standard rate for storage is $4m, annually
- The model has been deployed at the beginning of the Collider's Run-II in the Summer of 2015

CMS data certification / anomaly detection

Case



80%

saving on manual work on data certification tasks

Task

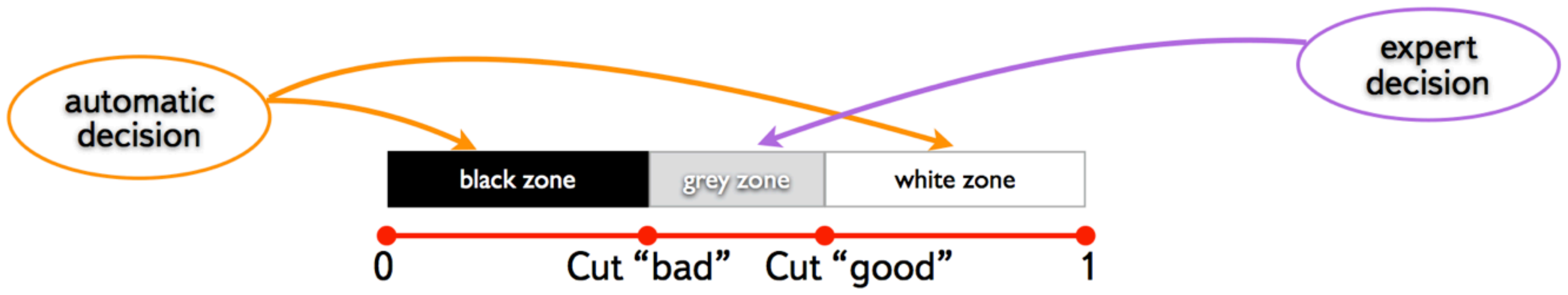
Traditionally, quality of the data at CMS experiment is determined manually. It requires considerable amount of human efforts;

Data used

- CERN open data portal 2010;
- Features: Particle flow jets, Calorimeter Jets, Photons, Muons;
- The dataset was labeled by CMS experts (~3 FTEs).

ML Metrics

- ROC AUC, precision

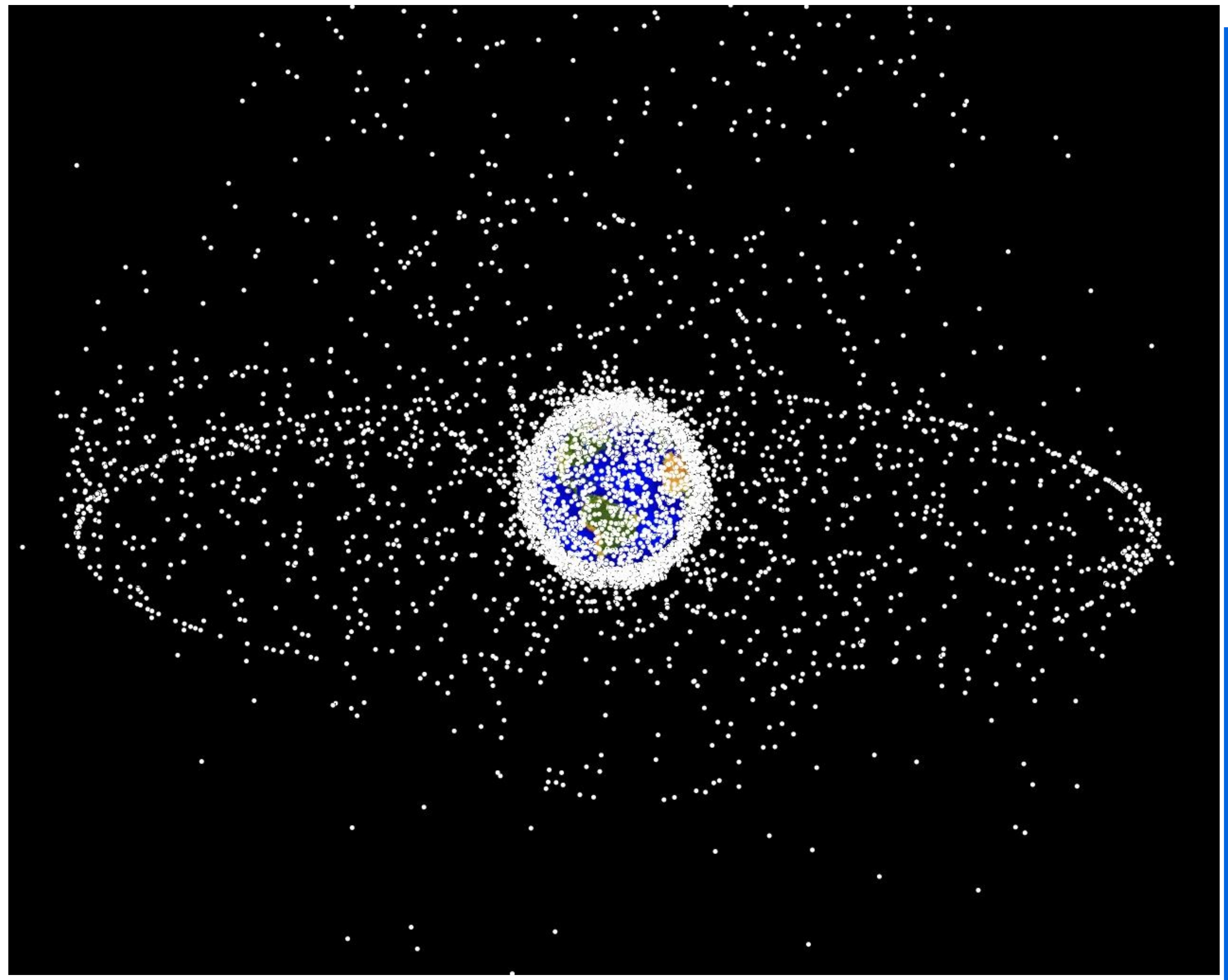


Result

- ~80% saving on manual work is feasible for Pollution & Loss rate of 0.5%.
- Next steps: adopt technique for 2016 data & run in production
- http://bit.ly/2l0MLiN

Satellite Control : Collision Avoidance

Case



https://en.wikipedia.org/wiki/Space_debris

$2 \cdot 10^{-4}$ %

probability of collision for 99% cases

Task

Reaction to a collision threat must be automated for large constellations (e. g. 10k planned in Starlink)

Decision is not a simple optimisation problem, it must consider many constraints

Methods

reinforced learning

Test

- 100 simulated conjunctions on Low Earth Orbit

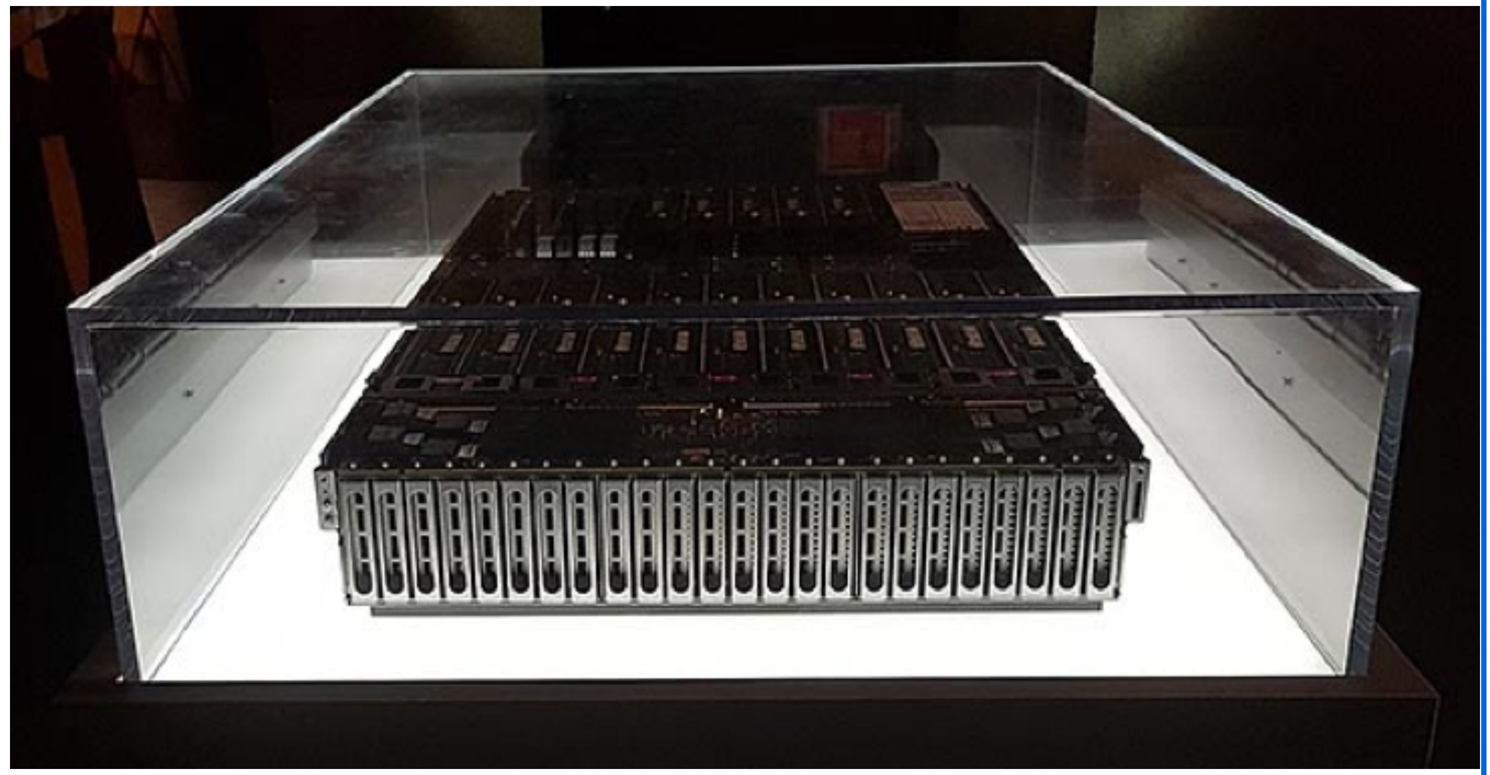
Result

- In the majority (68%) of cases SpaceNav fulfils all the constraints
- In almost every case (99%) it reduces the total collision probability to $2 \cdot 10^{-4}$
- The algorithm was configured to save fuel, by relaxing this requirement, any risk level can be achieved

arXiv:1902.02095

Satellite Control : Collision Avoidance

Case



https://en.wikipedia.org/wiki/Space_debris

99 %

failure detection rate, while false alarm rate is as low as 5%

Task

Anomalies and failures detection algorithm for early recover of the system

Construction of Digital Twin of Storage System for generation of rare events

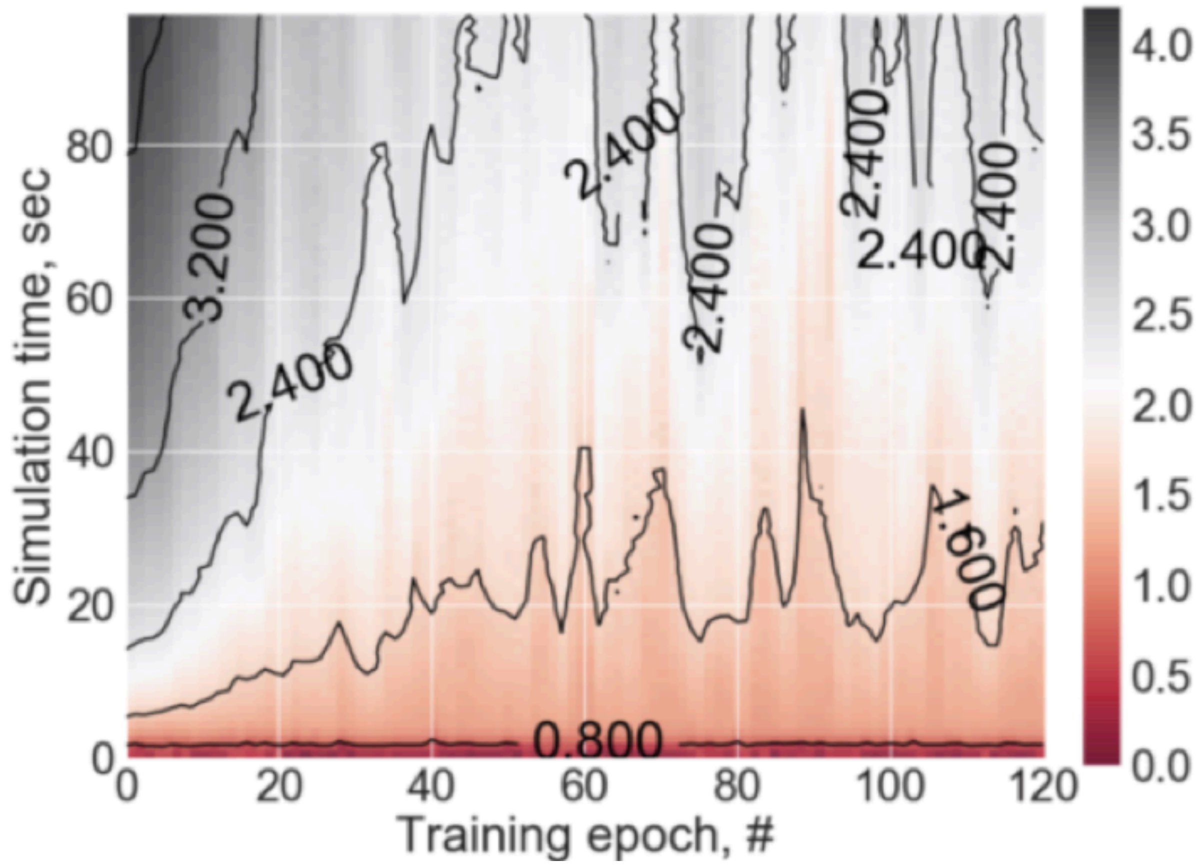
Methods

Proposed algorithm for anomaly detections for timeseries and implemented it in production-ready library: “Nostradamus”

Proposed combination of event-driven simulator and reinforcement learning control for simulator fine-tuning: “DeepController”

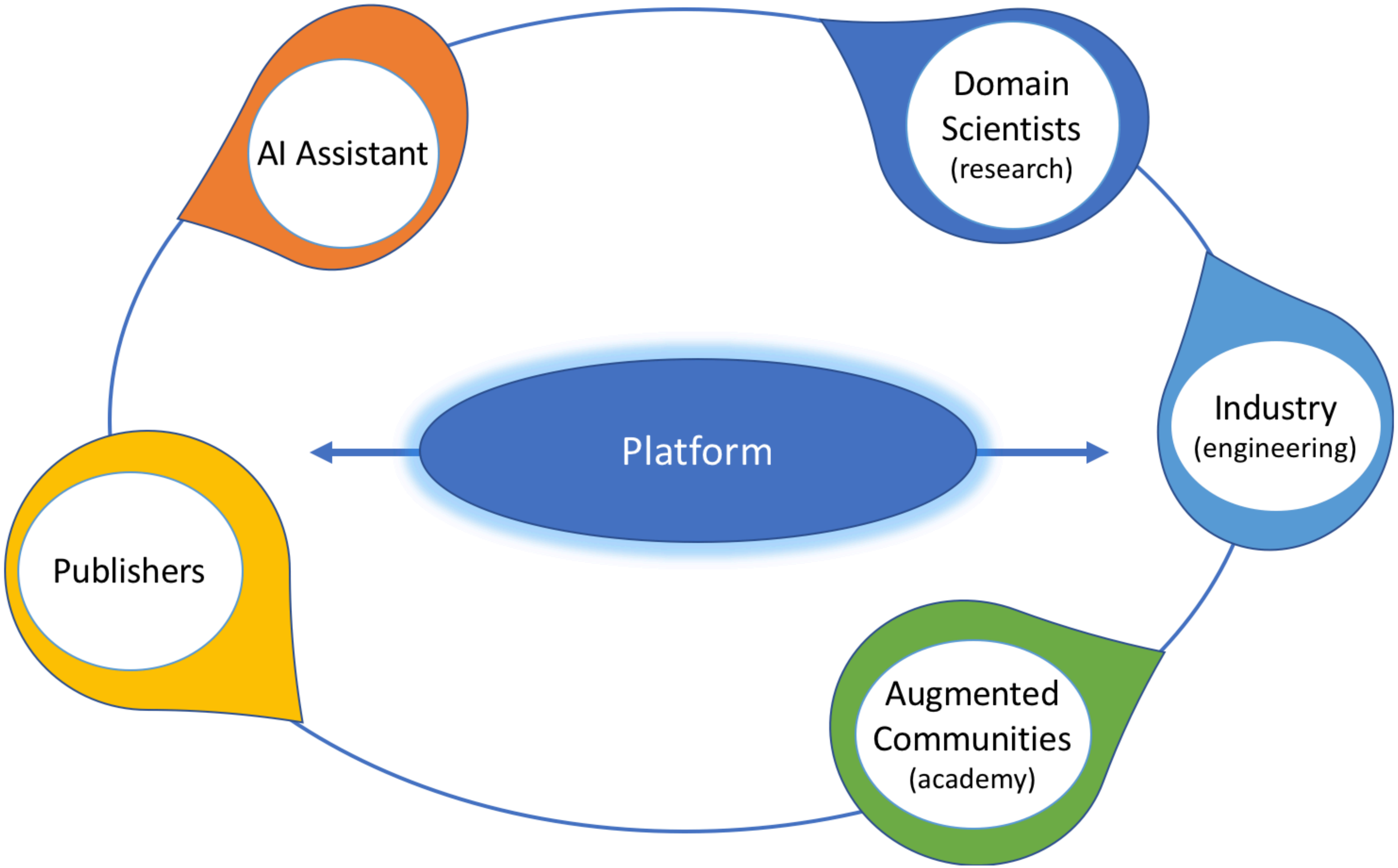
Result

- Failure Detection Rate is up to 99% while keeping False Alarm Rate at 5%
- After ~100 epochs quality of simulation improves in several times on target metrics in comparison with simulator without RL-control



Co-research methodology

Co-research platform



Co-Research Platform overview

Co-research competition platform

AI

Researchers

- Accelerate research by ML
- Solve challenges

Students

- Accelerate practical skills
- Develop own educational trajectory
- Learn to collaborate

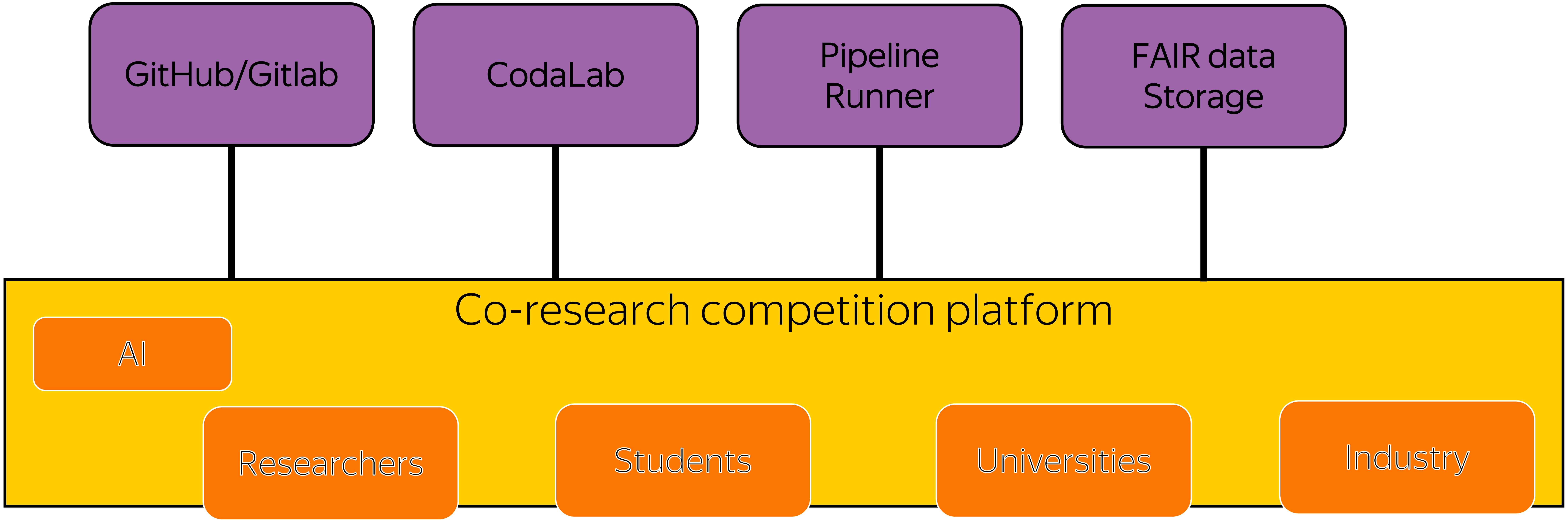
Universities

- Evaluate student skills
- Understand demand for new courses

Industry

- Get qualified employees
- Solve challenges

Co-Research Platform overview



https://coopetition.coresearch.club

Our collaborators

Particle physics

- LHCb (CERN), SHiP (CERN), CMS (CERN), OPERA (INFN), NewsDM (INFN)

Astrophysics

- http://www.sai.msu.ru/
- Institut für Astronomie und Astrophysik Tübingen

Neuroinformatics, Institute of Cognitive Neuroscience

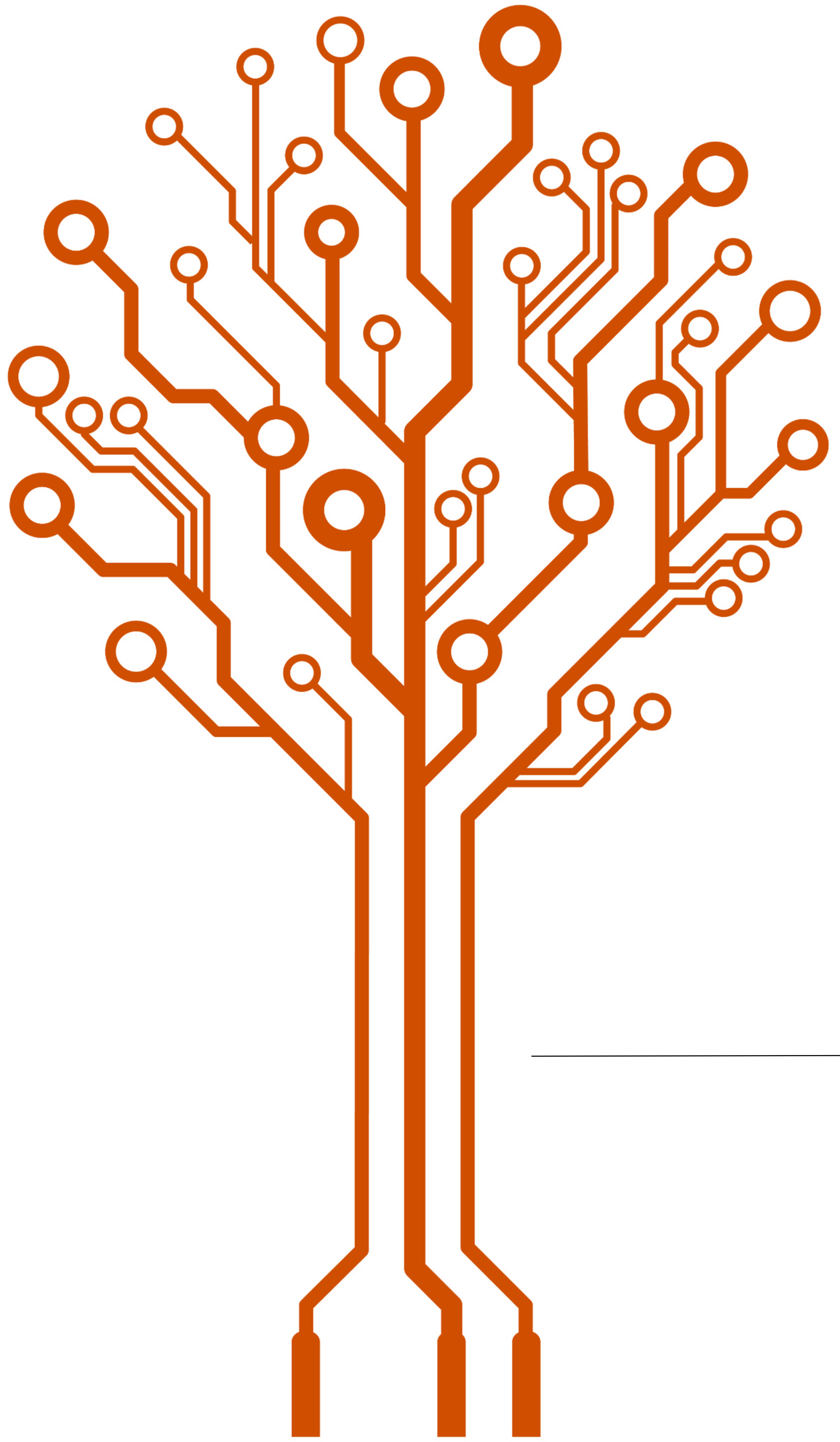
Space industry, Roscosmos

Metal production industry, MMK

Summary

- LambdaLab is focused on adopting advanced ML methods to challenging problems in Natural Science and Industry
- Many more completed and ongoing projects are not included in this presentation
- Educational efforts include regular lectures, courses, schools
- Staffed by 3 senior scientists, 6 researchers, 10 PhD students, MS students, …
- Close cooperation with teams in France, Germany, Italy, Switzerland, UK
- 30+ publications on applied ML for last 3 years

Thank you



Innovate with LAMBDA!

2020	
Head count	35
Applied Projects	9
Research Projects	30+

2015	
Head count	12
Applied Projects	2
Research Projects	10

hse_lambda
austyuzhanin@hse.ru

Backup

Recent research highlights

Advanced anomaly detection methods

Models capable of training on mixture of real and simulated data

Advanced generative models, digital twins

Differentiable surrogate models